

Learning Better Representations for Automated Speech Recognition

Dr. Shalini Ghosh

Principal Research Scientist, Amazon Alexa AI

Email: shalini.ghosh@gmail.com

Website: www.shalinighosh.com

Background

- **Core research focus:**
 - Multimodal Machine Learning (ML) modeling
 - Deep Learning for audio, vision and language
- **Previous work:**
 - Applying ML to dependable systems, e.g., cyber-security, fault-tolerance
- **Background:**
 - Director of ML Research @ Samsung Research America, 2+ years
 - Visiting Scientist @ Google Research, 1+ years
 - Principal Scientist @ Computer Science Lab, SRI International, 13+ years

More details and publication list at: shalinighosh.com

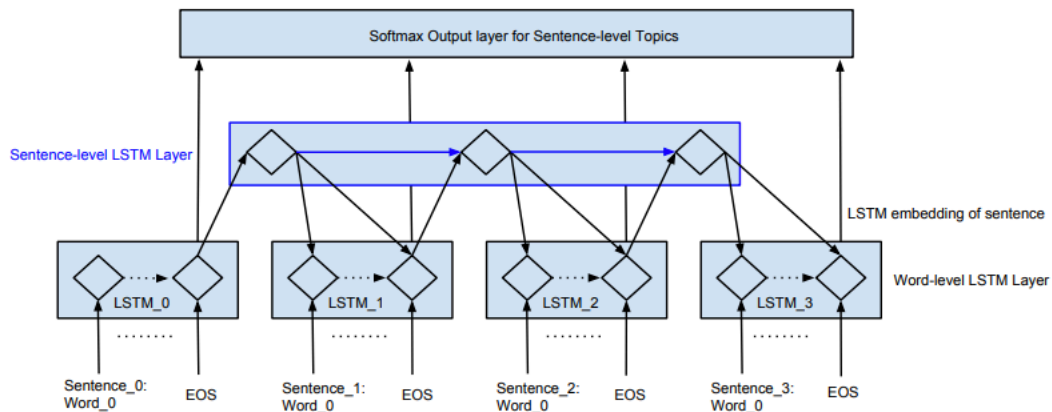
Overview of Talk

- Background
- How multimodal pre-training gives better representations for ASR
- How to get factorized content + context representations for ASR
- Key takeaways and Future work

Have worked on different aspects of **Multimodal representation learning**

Visiting Scientist @ Google Research

- Improve text representation of LSTM using shared contextual information (e.g., sentence topic) for next word/sentence prediction task [[KDD-DLW 2016](#)]



Have worked on different aspects of **Multimodal representation learning**

Principal Scientist @ SRI International

- Align text + image representations to come up with visual + textual explanations for Visual Question Answering task [[XAIW-ICML 2018](#)]

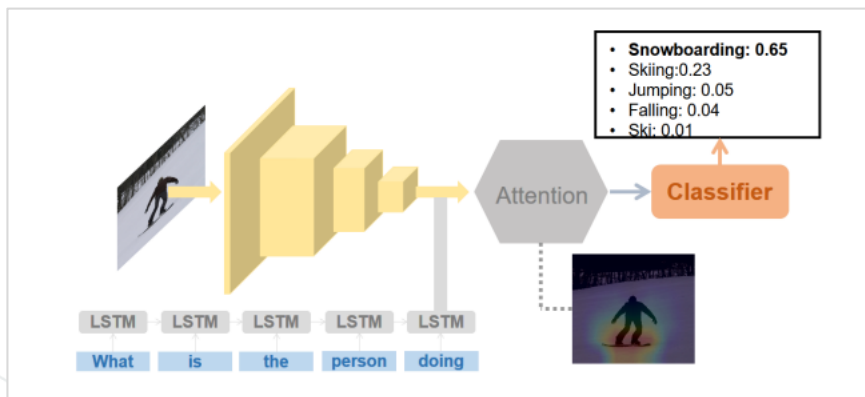
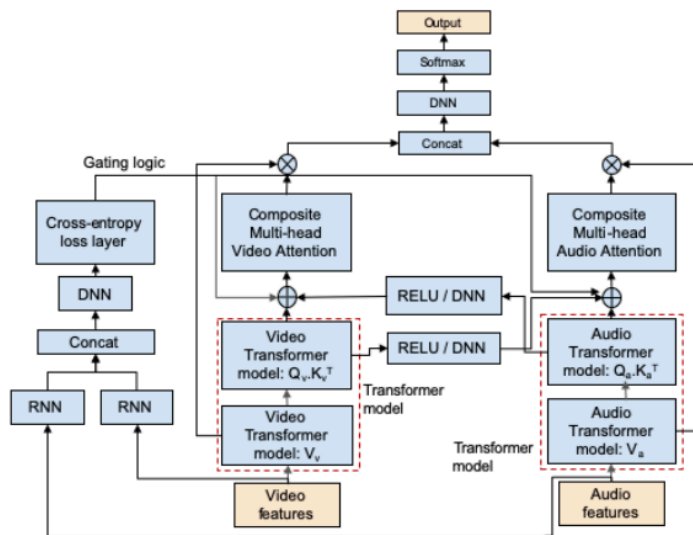


Figure 2: Salient parts of an Visual Genome image highlighted by the attention layers of our VQA Model, corresponding to the question/answer pair “What is this game? Tennis”.

Have worked on different aspects of **Multimodal representation learning**

Director @ Samsung Research America

- Use coherence (temporal + spatial) and correlation between modalities to improve visual & audio representation for video categorization task [[ACM-MM 2020](#)]



This talk: Better Representations for ASR in Alexa Speech

Multimodal Pre-Training for Automated Speech Recognition

David M. Chan, Shalini Ghosh, Debmalya Chakrabarty, Björn Hoffmeister

2022 IEEE International Conference on Acoustics, Speech and Signal Processing

Combine globally robust representations with locally accurate representations

Content-Context Factorized Representations for Automated Speech Recognition

David M. Chan, Shalini Ghosh

(Submitted) INTERSPEECH 2022

Decoupling content representations from background noise

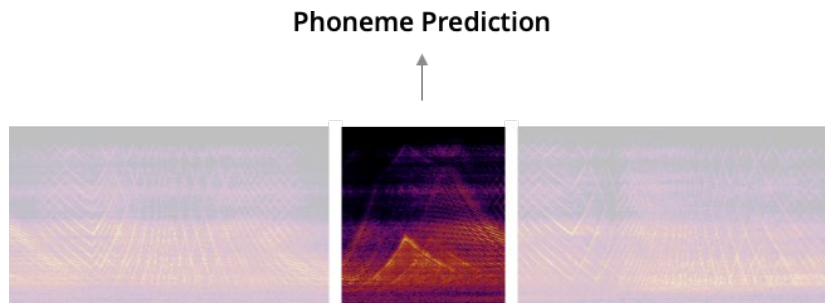
Many ASR models focus on extracting phonemes from local waveform representations

Local first is good:

- Models are less likely to hallucinate
- Models focus on “what matters”

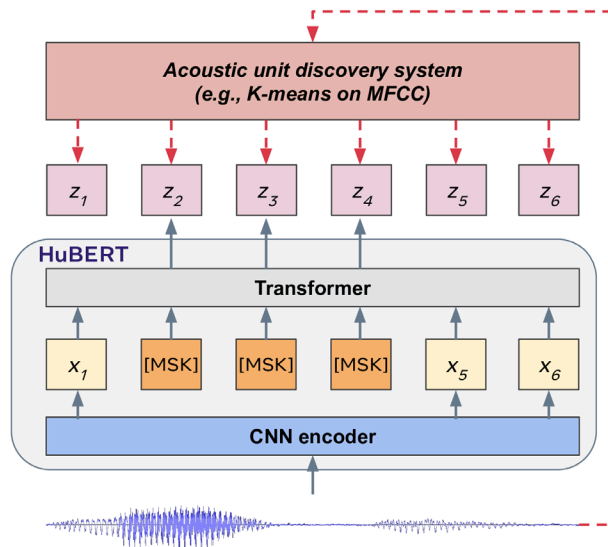
But...

- Local-first models are not context-aware
- They struggle to disambiguate phonemes when there is global noise or signal aberrations



Recent transformer-based architectures (e.g., HuBERT) allow models to incorporate global context

- Through masked reconstruction targets
- Using self-supervised learning

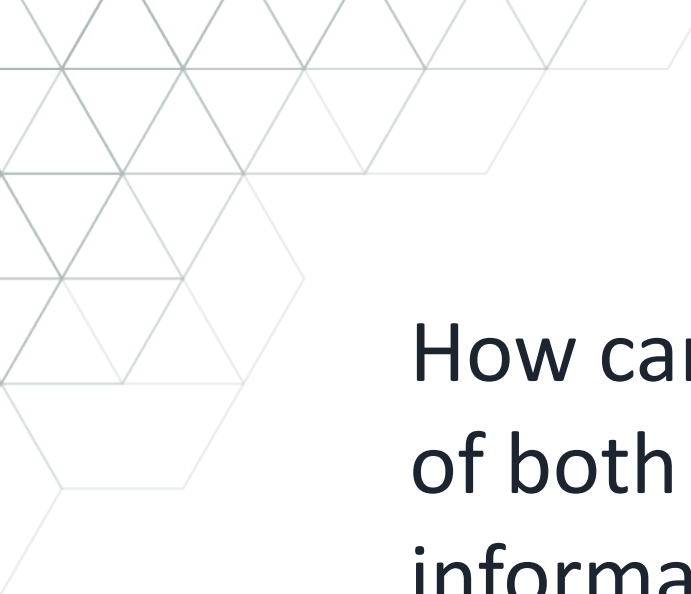


Global context is good:

- Models tend to be robust to noise events
- Large amounts of pre-training can help models build strong statistical priors for unseen data scenarios

But...

- Models can be prone to hallucination, and will often trust priors over local evidence
- Models often tend to encode low-frequency info over high-frequency signal, due to loss targets



How can we leverage the strengths
of both global and local
information in learning audio
representations?

How can we build a good global context embedding?

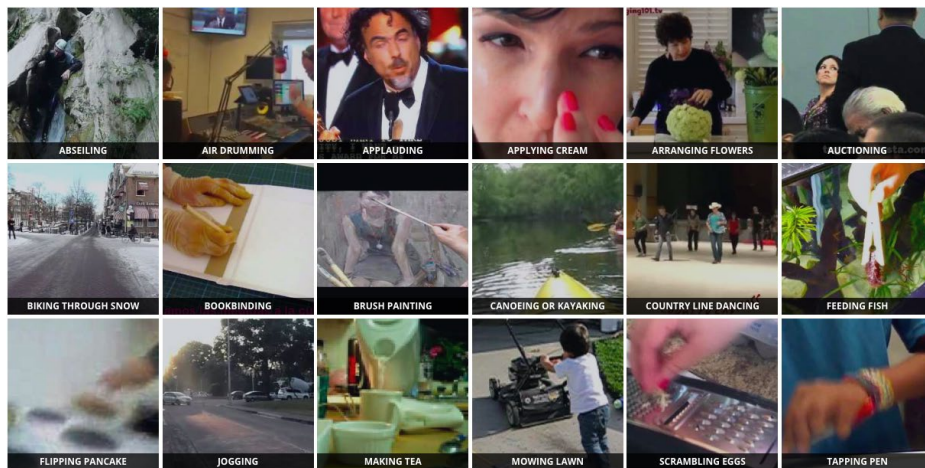
#1 Cover a wide range of scenarios and auditory environments

#2 Encode semantic information as much as possible

#3 Generate representations that give useful contextual information

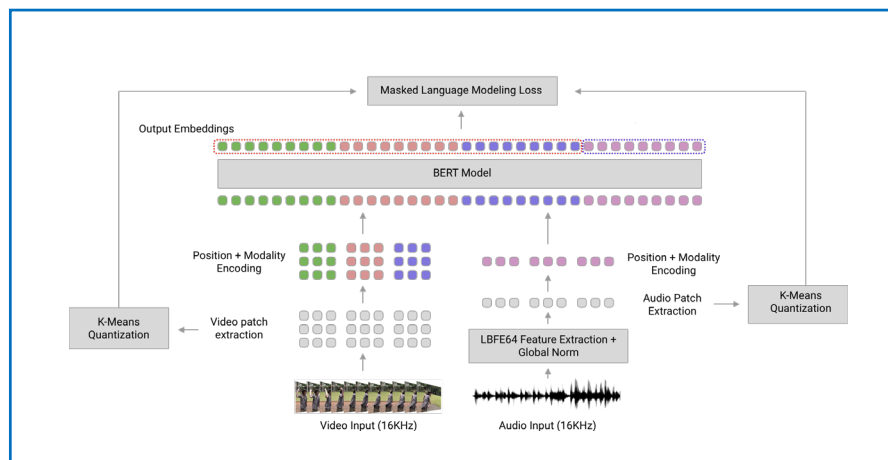
Building a good global context embedding from videos

- Audio-only data from several sources can serve as a powerful way to augment representations
- **Video data** contains such varied audio data, but it also contains additional **visual context**
- **Visual context** can serve as pseudo concept-labeled, and improve audio embedding spaces by linking traditionally uncorrelated auditory events.

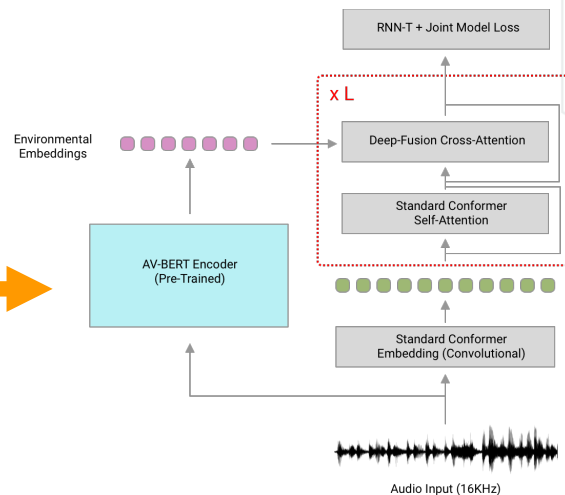


Incorporating pre-trained AV-BERT into ASR model

Stage 1



Stage 2



- Fuse a standard Conformer-based audio encoder with a pre-trained **audio/video BERT** model
- Leverage cross-attention to bring context information into local waveform representations
- Use learned audio-video representations at test time, **even without the video**

Improvements on Librispeech

Method	Params (M)	test-clean	test-other
LAS			
Transformer [25]	370	2.89	6.98
Transformer [26]	-	2.2	5.6
LSTM [2]	360	2.6	6.0
Transducer			
Transformer [5]	139	2.4	5.6
ContextNet (M) [27]	31.4	2.4	5.4
ContextNet (L) [27]	112.7	2.1	4.6
Conformer			
Conformer (M) [2]	30.7	2.3	5.0
Conformer (L) [2]	118.8	2.1	4.3
Ours			
Conf. (M, base)	79	2.21	4.85
Conf. (L, base)	122	2.11	4.29
A + Conf. (M)	79	2.15 (+2.7%)	4.82 (+0.6%)
A/V + Conf. (M)	79	2.10 (+4.8%)	4.72 (+2.7%)
A/V + Conf. (L)	122	1.98 (+7.0%)	4.10 (+4.4%)

Improvements on Small Models

Method	Base	Rare	Query	Messages
Conformer (M)	0	0	0	0
+ Audio (M)	+30.1%	+17.9%	+26.7%	+20.1%
+ Audio/Video (M)	+45.6%	+31.2%	+38.7%	+17.2%
Conformer (L)	0	0	0	0
+ Audio/Video (L)	+5.1%	+5.4%	+4.2%	+5.9%

Qualitative Example

Reference	Should	I	buy	from	the	princess	***	***	starfrost	set	royale	high
Baseline	Should	I	buy	from	the	princess	sleaze	forest	in	we're	all	rarehide
Ours	Should	I	buy	from	the	princess	***	sleaz	frost	set	royale	high

Better Representation for Automated Speech Recognition

Multimodal Pre-Training for Automated Speech Recognition

David M. Chan, Shalini Ghosh, Debmalya Chakrabarty, Björn Hoffmeister
2022 IEEE International Conference on Acoustics, Speech and Signal Processing

Combine globally robust representations with locally accurate representations

Content-Context Factorized Representations for Automated Speech Recognition

David M. Chan, Shalini Ghosh
(Submitted) INTERSPEECH 2022

Decoupling content representations from background noise

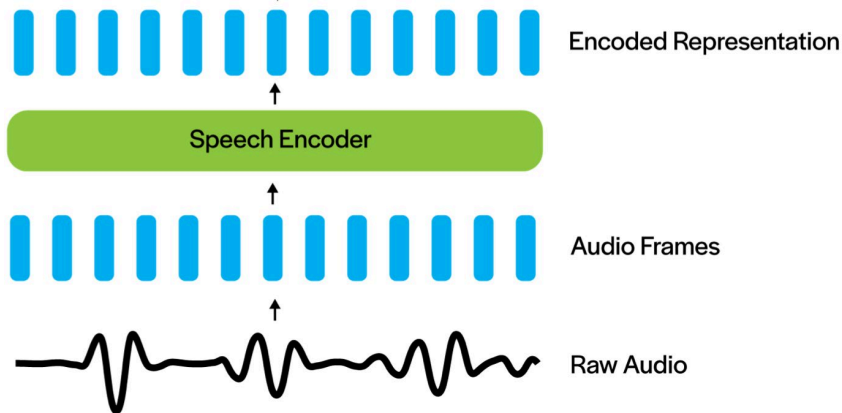
Many ASR models focus on encoding content + context in unified representations

Unified encoding is good:

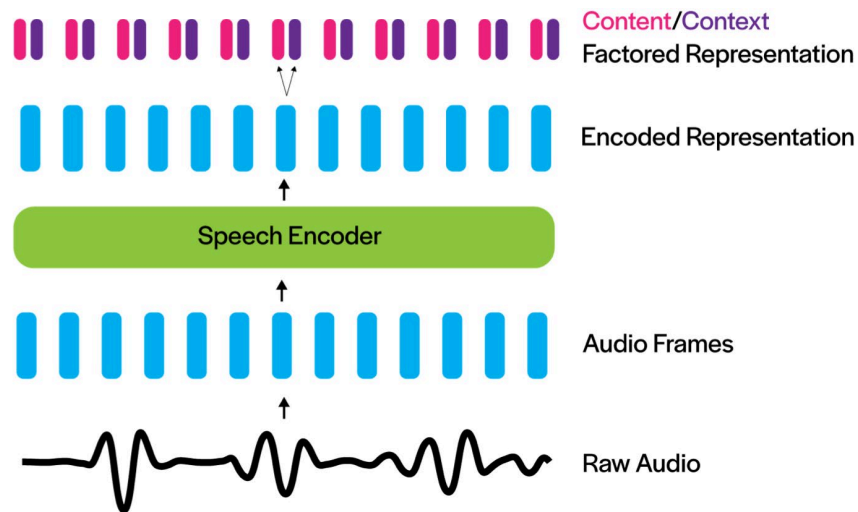
- Unified encoders have simple training architectures
- Unified encoders can make use of all information to achieve a targeted downstream task
- Unified encoders require little downstream user intervention

But...

- Confusion between information relevant for a task and irrelevant information can encourage models to rely on spurious correlations in the data
- Such spurious correlations are made worse in the long tails of ASR distributions, and in multi-modal models such as AV-BERT



Our proposed model introduces context-content factorization, to explicitly separate understanding “what” from understanding “where”



How can we effectively separate content and context?

1) Maximize mutual information between the content representation and the correct ASR labels

$$\max_{\theta} I(\phi_{\text{content}}(x, \theta); y)$$

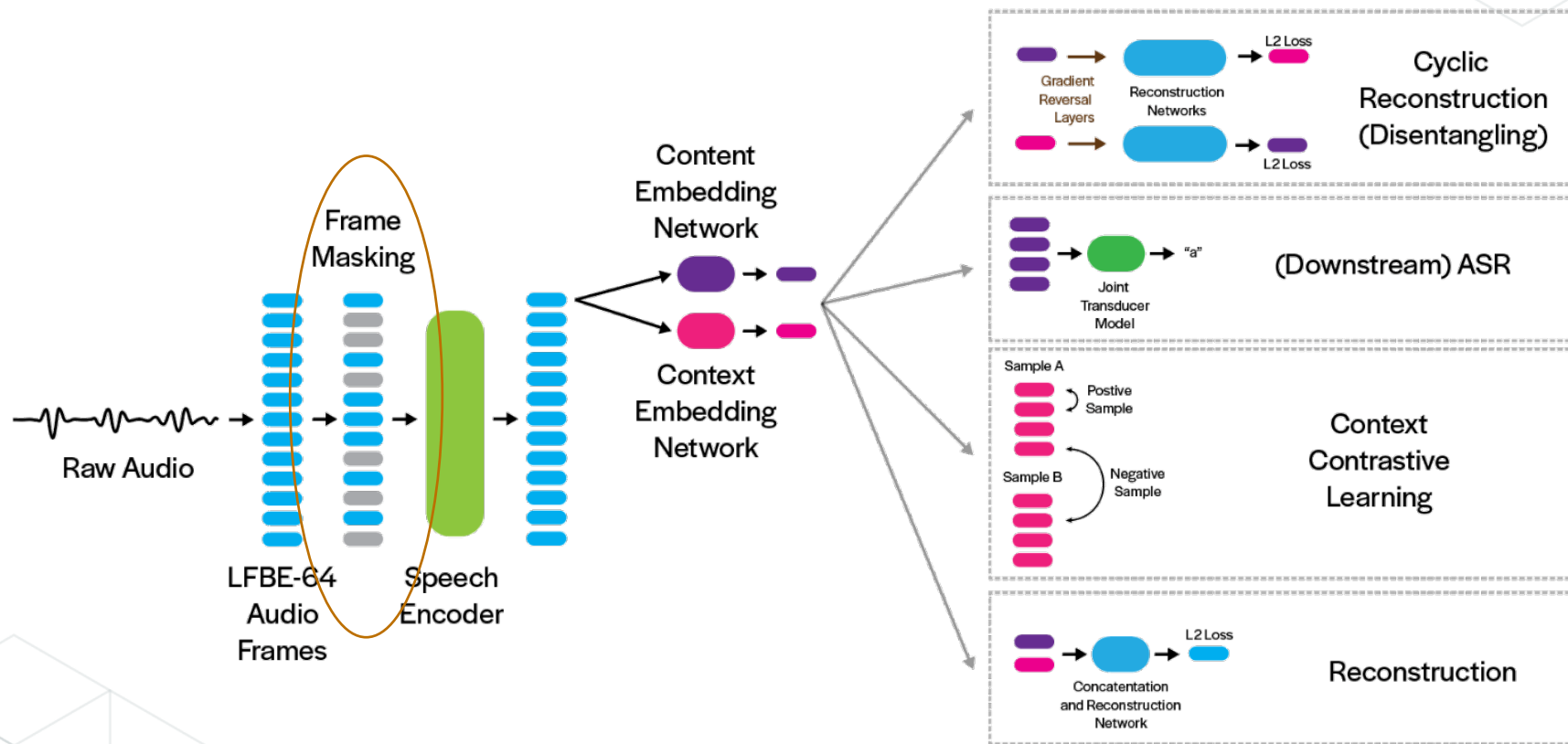
2) Encourage the content representation to encode background features, noise, and speaker identity

$$\max_{\theta} I(\phi_{\text{context}}(x, \theta); P_{x|y})$$

3) Minimize mutual information between content representation and context representation

$$\min_{\theta} I(\phi_{\text{context}}(x, \theta); \phi_{\text{content}}(x, \theta))$$

Proposed Model: Factored Content-Context Representations for ASR



Improvements on Librispeech

- Without the factored representation, only MLM shows performance gains over the baseline models.
- Introducing factored representations improves generalization in both RNN-t and Conformer models
- Using all losses is useful: and more useful in the test-other dataset, where noise is more likely to be prevalent (and the model needs to rely on factored information)

Method	FR	BG-C	MLM	test-clean	test-other
RNN-T					
			<i>Baseline</i>	6.05	15.43
	✓			6.06 (+0.01%)	15.42 (-0.06%)
		✓		6.08 (+0.05%)	15.46 (+0.19%)
			✓	5.98 (-1.16%)	15.07 (-2.33%)
	✓	✓		5.91 (-2.31%)	15.01 (-2.72%)
	✓		✓	5.87 (-2.97%)	14.84 (-3.82%)
	✓	✓	✓	5.80 (-4.13%)	14.61 (-5.31%)
Conformer					
			<i>Baseline</i>	2.13	4.31
	✓			2.12 (-0.47%)	4.28 (-0.70%)
		✓		2.14 (+0.47%)	4.30 (-0.23%)
			✓	2.10 (-1.41%)	4.26 (-1.16%)
	✓	✓		2.11 (-0.90%)	4.20 (-2.55%)
	✓		✓	2.09 (-1.87%)	4.16 (-3.48%)
	✓	✓	✓	2.07 (-2.81%)	4.10 (-4.77%)

Key Take-aways

Self-supervised learning can be used to augment current local techniques with global representations

Multi-modal representations can help, even if only one mode is available at test time

Factoring content and context can help us to better model clean and noisy ASR data

Leveraging **auxiliary losses** for better factored learning can help lead to a more holistic model of ASR data

Future work

Find new sources of **context** to leverage for further ASR performance improvement, e.g., multilingual data, different environments

Explore **contextual** and situational information to improve the models and make better ASR decisions

Going beyond ASR, understand how multi-modal context information can be leveraged in **end-to-end dialog understanding** tasks (including NLU)